

## CLAIMS

1           1. A method for identifying sequences of molecules and sequence modifications from mass  
2 spectrometry data comprising:  
3           a. producing at least one *de novo* sequence from mass spectrometry data of sequences of  
4 molecules,  
5           b. calculating at least one mass-based alignment between each *de novo* sequence and sequences  
6 in a sequence database, wherein the molecular masses of molecules in the *de novo* sequence are compared  
7 to molecular masses of molecules in each sequence in the sequence database,  
8           c. interpreting mass differences of modification sites between the sequence in the sequence  
9 database and the *de novo* sequence that have been identified by the mass-based alignment as  
10 modifications identified in a modification catalog,  
11           d. calculating at least one match score for the mass-based alignment that provides an indication  
12 of matching between the sequence in the sequence database and the *de novo* sequence,  
13           e. identifying sequences in the sequence database from mass-based alignments in response to the  
14 match scores, and  
15           f. grouping identifications of sequences in the sequence database from at least one *de novo*  
16 sequence into an identified macromolecule list that agrees with the *de novo* sequencing results.

1           2. The method of claim 1, wherein the mass spectrometry data is generated from a tandem mass  
2 spectrometer device.

1           3. The method of claim 1, wherein at least one *de novo* sequence is an estimated sequence of  
2 molecules generated from the mass spectrometry data derived from a sequence of molecules.

1           4. The method of claim 3, wherein a *de novo* sequence is a complete or partial sequence of  
2 molecules.

1           5. The method of claim 3, wherein a *de novo* sequence contains incorrect or unidentifiable region  
2 of molecules where the exact sequence of molecules cannot be determined.

1           6. The method of claim 5, wherein a mass region is the molecular mass of the molecules in an  
2 unidentifiable region of molecules.

1           7. The method of claim 1, wherein at least one molecule is an amino acid and at least one  
2 sequence of molecules is a peptide.

- 1           8. The method of claim 7, wherein the peptides are derived by an enzymatic digestion of  
2 proteins.
- 1           9. The method of claim 7, wherein the sequence database is a database of amino acid sequences  
2 of proteins.
- 1           10. The method of claim 7, wherein the sequence database is a database of amino acid sequences  
2 derived from nucleotide sequences.
- 1           11. The method of claim 7, wherein the sequence database is a database of *de novo* peptide  
2 sequences.
- 1           12. The method of claim 7, wherein the sequence in the sequence database is a particular amino  
2 acid sequence in the sequence database.
- 1           13. The method of claim 6, further comprising:  
2           a. identifying a sequence in the sequence database with a tag match, and  
3           b. generating a mass-based alignment between a *de novo* sequence and the sequence in the  
4 sequence database.
- 1           14. The method of claim 13, wherein a mass-based alignment is a series of consecutive local  
2 mass-based alignments on either side of a tag match.
- 1           15. The method of claim 14, wherein a tag match is when a tag in the *de novo* sequence has been  
2 shown to be equivalent to a tag in a sequence in the sequence database by way of a tag search.
- 1           16. The method of claim 15, wherein a tag search is used to identify a subset of sequences in the  
2 sequence database from which to compute mass-based alignments.
- 1           17. The method of claim 16, wherein a tag is a sequence of consecutive molecules of a specified  
2 length, and the specified length is 2 to 4 molecules in length.
- 1           18. The method of claim 16, wherein single molecules of the tag and sequences in the sequence  
2 database that have the same nominal weight are represented by a single molecule.
- 1           19. The method of claim 14, wherein molecules at either side of the tag match in both the *de*  
2 *novo* sequence and the sequence of the sequence database are converted into mass objects.

1           20. The method of claim 19, wherein a mass object is at least one molecular mass and a name for  
2 that mass.

1           21. The method of claim 18, wherein for single molecules, mass objects are assigned the  
2 molecular mass of the single molecule.

1           22. The method of claim 18, wherein for unidentifiable mass regions, mass objects are  
2 assigned the molecular mass of the unidentifiable mass region.

1           23. The method of claim 18, wherein for reference amino acids, which represent multiple  
2 amino acids, mass objects are assigned the molecular mass of each amino acid.

1           24. The method of claim 19, wherein for variably modified amino acids, mass objects are  
2 assigned multiple molecular masses.

1           25. The method of claim 19, wherein mass regions are treated as single molecules with a single  
2 molecular mass.

1           26. The method of claim 19, wherein a gap is a mass object of zero molecular mass that  
2 represents no molecule.

1           27. The method of claim 19, wherein a local mass-based alignment is a matching of at least one  
2 consecutive mass object in the sequence in the sequence database and at least one consecutive mass object  
3 in a *de novo* sequence.

1           28. The method of claim 27, wherein each local mass-based alignment is generated with a  
2 breadth-first search, wherein all possible sequential combinations of mass objects of the next specified  
3 number of mass objects are compared.

1           29. The method of claim 28, wherein the specified number of mass objects used in the breadth  
2 first search is the search depth.

1           30. The method of claim 29, wherein the search depth is 3-5.

1           31. The method of claim 21, wherein the breadth first search is used identify the local mass-  
2 based alignment as either a mass match, a substitution, or a gap match:

1           32. The method of claim 31, wherein the breadth first search first tries to identify a mass match,  
2 as a local mass-based alignment where the sum of the molecular masses of the consecutive mass objects

3 in the sequence in the sequence database and the sum of the molecular masses of the consecutive mass  
4 objects in a *de novo* sequence are equal within a specified mass tolerance.

1 33. The method of claim 31, wherein if there are no mass objects left on the side of the tag match  
2 in the sequence in the sequence database, a gap match is identified as a local mass-based alignment  
3 between a gap and at least one consecutive mass object in either the sequence in the sequence database or  
4 the *de novo* sequence.

1 34. The method of claim 31, wherein if a mass match or a gap cannot be identified, then the  
2 breadth first search identifies a modification site as a local mass-based alignment where the sum of the  
3 molecular masses of the consecutive mass objects in the sequence in the sequence database and the sum  
4 of the molecular masses of the consecutive mass objects in a *de novo* sequence are not equal within a  
5 specified mass tolerance.

1 35. The method of claim 31, wherein the number of mass objects in the *de novo* sequence and the  
2 number of mass objects in the sequence database is minimized.

1 36. The method of claim 31, wherein the specified mass tolerance is designated by a mass  
2 tolerance of a tandem mass spectrometer device that generates the mass spectrometry data.

1 37. The method of claim 28, wherein a new local mass-based alignment is generated starting  
2 from the next molecule in the *de novo* sequence and the next molecule in the sequence in the sequence  
3 database after the last molecule that is matched in the breadth-first search in each respective sequence.

1 38. The method of claim 37, wherein a series of local mass-based alignments are made until the  
2 entire *de novo* sequence has been accounted for by the sequence in the sequence database in the mass-  
3 based alignments.

1 39. The method of claim 38, wherein a maximum number of consecutive modification sites are  
2 performed.

1 40. The method of claim 39, wherein the maximum number of consecutive modification sites is  
2 1-or 2 local mass-based alignments in length.

1 41. The method of claim 39, wherein the modification information about modifications is  
2 cataloged in a modification catalog.

1 42. The method of claim 41, wherein the modification information includes at least one of,  
2 molecular mass of the modification, a specific molecules where the modification occurs, a frequency of

3 occurrence of the modification at those molecules, wherein the frequency of occurrence is the estimated  
4 frequency in nature or a frequency as a sample preparation artifact, a mass object for the modification,  
5 which represents the additional mass of the modification to the *de novo* sequence at those molecules, and  
6 the name of the modification, and a modification score for the modification.

1 43. The method of claim 42, wherein a modification is selected from, an *in vivo* or *in vitro*  
2 protein, a peptide modification, and an amino acid substitution.

1 44. The method of claim 43, further comprising: ranking the modifications, wherein the ranking  
2 is based on their frequency of occurrence.

1 45. The method of claim 44, further comprising: identifying a most probable modification in the  
2 modification site from the modification catalog by matching elements to elements in modifications in the  
3 modification catalog that are selected from at least one of, the mass difference, the molecules in the  
4 sequence database in the modification site, and the ranking of the modification in the modifications  
5 catalog.

1 46. The method of claim 45, wherein the mass difference is the difference between the sum of  
2 the molecular masses of the consecutive mass objects in the sequence in the sequence database and the  
3 sum of the molecular masses of the consecutive mass objects in a *de novo* sequence in a local mass-based  
4 alignment.

1 47. The method of claim 45, wherein the mass object of an identified modification is inserted  
2 into the in the mass-based alignment, which creates a mass match between the *de novo* sequence and the  
3 sequence in the sequence database.

1 48. The method of claim 38, further comprising: computing a match score of the mass-based  
2 alignment, the match score being a measure of how well the sequence in the sequence database matches  
3 the *de novo* sequence.

1 49. The method of claim 48, wherein a match score is generated from the linear combination of  
2 local alignment scores from the series of local mass-based alignments.

1 50. The method of claim 49, wherein each of a series of consecutive local mass-based alignments  
2 receives a score and is classified.

1 51. The method of claim 50, wherein each local alignment score is generated using a substitution  
2 matrix, depending on whether the local alignment is a mass match, a modification site, or a gap match.

1           52. The method of claim 51, wherein the substitution matrix contains substitution matrix score of  
2 least one molecule.

1           53. The method of claim 52, wherein the substitution matrix identity score is a substitution  
2 matrix score between a molecule and itself.

1           54. The method of claim 53, wherein the substitution matrix substitution score is a substitution  
2 matrix score between a molecule and a different molecule.

1           55. The method of claim 54, wherein the substitution matrix score is the log of the odds score of  
2 an identity of a molecule or a substitution between two molecules.

1           56. The method of claim 52, wherein the local alignment score for a mass match is the average  
2 value of the substitution matrix identity scores for all of the molecules in the sequence in the sequence  
3 database matched in the local alignment.

1           57. The method of claim 56, wherein if at least one of the molecules has been modified by a  
2 modification, the substitution matrix score for each modified molecule is the modification score for that  
3 modification.

1           58. The method of claim 52, wherein if the local mass-based alignment is a match between only  
2 one mass object from the sequence in the sequence database, and only one mass object from the *de novo*  
3 sequence, and that those mass objects represent single molecules, then the local alignment score for a  
4 substitution is the substitution matrix substitution score between the molecule in the sequence in the  
5 sequence database and the molecule in the *de novo* sequence.

1           59. The method of claim 52, wherein the local alignment score for a substitution is the number of  
2 molecules in the substitution in the sequence in the sequence database multiplied by the average value of  
3 the substitution matrix substitution scores.

1           60. The method of claim 52, wherein the local alignment score for a gap match is the number of  
2 molecules in the gap match in the *de novo* sequence multiplied by the average value of the substitution  
3 matrix substitution scores.

1           61. The method of claim 48, wherein if the termini of the *de novo* sequence are expected to be  
2 specific molecules, the match score is increased if the termini of the mass-based alignment match the  
3 expected specific molecules.

1           62. The method of claim 48, wherein if the termini of the *de novo* sequence are expected to be  
2 specific molecules, the match score is decreased if the termini of the mass-based alignment do not match  
3 the expected specific molecules, or if expected specific molecules are present inside the mass-based  
4 alignment.

1           63. The method of claim 1, further comprising utilizing an approach that interprets matches  
2 between sequences in the sequence database and *de novo* sequences, which have been scored by a match  
3 score, as an identified macromolecule list and assigns a macromolecule score to each sequence in the  
4 identified macromolecule list.

1           64. The method of claim 63, wherein the match score is a measure of how well the sequence in  
2 the sequence database matches the *de novo* sequence.

1           65. The method of claim 64, wherein *de novo* sequences that match at least one sequence in the  
2 sequence database are classified as either discriminating *de novo* sequences or non-discriminating *de novo*  
3 sequences, the *de novo* sequences are inserted into a *de novo* sequence list, and the *de novo* sequences in  
4 the *de novo* sequence list are ranked by their delta scores.

1           66. The method of claim 65, wherein the delta score is computed for the *de novo* sequence as the  
2 difference between the match scores of the first and second matches to sequences in the sequence  
3 database for that *de novo* sequence. If that *de novo* sequence only matches one sequence in the sequence  
4 database, the delta score is the match score for that match.

1           67. The method of claim 66, wherein discriminating *de novo* sequences have a delta score greater  
2 than or equal to the delta score threshold and non-discriminating *de novo* sequences have a delta score  
3 less than the delta score threshold.

1           68. The method of claim 67, wherein the delta score threshold for the *de novo* sequence is  
2 between 0% and 25% of the match score of the highest scoring match between a sequence in the sequence  
3 database and that *de novo* sequence.

1           69. The method of claim 67, All matches between a sequence in the sequence database and a *de*  
2 *novo* sequence with match scores less than the match score of the highest scoring match between a  
3 sequence in the sequence database and that *de novo* sequence minus the delta score threshold are  
4 discarded.

1           70. The method of claim 60, wherein the sequence in the sequence database, which matches best  
2 to the discriminating *de novo* sequence in the *de novo* sequence list with the greatest delta score, is added  
3 to the identified macromolecule list. This *de novo* sequence is then moved from the *de novo* sequence list  
4 to that sequence.

1           71. The method of claim 70, wherein all non-discriminating *de novo* sequences in the *de novo*  
2 sequence list that match to that sequence in the identified macromolecule list are moved from the *de novo*  
3 sequence list to that sequence.

1           72. The method of claim 71, wherein the process of 1 is repeated until all discriminating *de novo*  
2 sequences in the *de novo* sequence list are removed from the *de novo* sequence list.

1           73. The method of claim 72, wherein all sequences in the sequence database that match to non-  
2 discriminating *de novo* sequences still in the *de novo* sequence list are added to the identified  
3 macromolecule list, and the non-discriminating *de novo* sequences still in the *de novo* sequence list are  
4 moved to those sequences.

1           74. The method of claim 73, wherein a macromolecule score is generated for every sequence in  
2 the identified macromolecule list.

1           75. The method of claim 74, wherein the macromolecule score is a linear combination of the *de*  
2 *novo* macromolecule scores of the *de novo* sequences that have been assigned to that sequence.

1           76. The method of claim 64, a new sequence database is generated containing only the sequences  
2 in the sequence database that are listed in the identified macromolecule list.

1           77. The method of claim 76, wherein *de novo* sequences that do not match any sequence in the  
2 original sequence database are re-analyzed by calculating a mass-based alignment between each *de novo*  
3 sequence in question and sequences in the new sequence database, as described in claim 1 in a way that  
4 the search space explored by the mass-based alignment algorithm is increased.

1           78. The method of claim 77, further comprising: decreasing the specified length of tags.

1           79. The method of claim 77, further comprising: increasing the search depth.

1           80. The method of claim 77, further comprising: increasing the maximum number of consecutive  
2 substitutions.



1           81. The method of claim 64, wherein *de novo* sequences that do not match any sequence in the  
2 original sequence database are re-analyzed by calculating a mass-based alignment between each *de novo*  
3 sequence in question and sequences in a different sequence database, as described in claim 1.

1           82. A method for identifying sequences of molecules and sequence modifications from mass  
2 spectrometry data comprising:  
3           a. producing at least one *de novo* sequence from mass spectrometry data of sequences of  
4 molecules,  
5           b. calculating at least one mass-based alignment between each *de novo* sequence and sequences  
6 in a sequence database, wherein the molecular masses of molecules in the *de novo* sequence are compared  
7 to molecular masses of molecules in each sequence in the sequence database,  
8           c. interpreting mass differences of modification sites between the sequence in the sequence  
9 database and the *de novo* sequence that have been identified by the mass-based alignment as  
10 modifications identified in a modification catalog, and  
11           d. calculating at least one match score for the mass-based alignment that provides an indication  
12 of matching between the sequence in the sequence database and the *de novo* sequence.

1           83. The method of claim 82, further comprising: identifying sequences in the sequence database  
2 from mass-based alignments in response to the match scores.

1           84. The method of claim 83, further comprising: grouping identifications of sequences in the  
2 sequence database from at least one *de novo* sequence into an identified macromolecule list that agrees  
3 with the *de novo* sequencing results.

1           85. A computer readable medium having stored thereon instructions which, when executed by a  
2 processor, cause the processor to perform:  
3           a. executing a first application that produces at least one *de novo* sequence from mass  
4 spectrometry data of sequences of molecules,  
5           b. executing a second application that calculates at least one mass-based alignment between each  
6 *de novo* sequence and sequences in a sequence database, wherein the molecular masses of molecules in  
7 the *de novo* sequence are compared to molecular masses of molecules in each sequence in the sequence  
8 database,  
9           c. executes a third program that interprets mass differences of modification sites between the  
10 sequence in the sequence database and the *de novo* sequence that have been identified by the mass-based  
11 alignment as modifications identified in a modification catalog, and

12 d. executes a fourth program that calculates at least one match score for the mass-based  
13 alignment that provides an indication of matching between the sequence in the sequence database and the  
14 *de novo* sequence.

1 86. The computer readable medium of claim 85, wherein the processor further executes a fifth  
2 program that identifies sequences in the sequence database from mass-based alignments in response to the  
3 match scores.

1 87. The computer readable medium of claim 86, wherein the processor further executes a sixth  
2 program that groups identifications of sequences in the sequence database from at least one *de novo*  
3 sequence into an identified macromolecule list that agrees with the *de novo* sequencing results.

1 88. A computer based system that implements identification sequences of molecules and  
2 sequence modifications from mass spectrometry data, comprising at least a first processor that executes  
3 one or more programs that:

4 a. produces at least one *de novo* sequence from mass spectrometry data of sequences of  
5 molecules,

6 b. executing a second application that calculates at least one mass-based alignment between each  
7 *de novo* sequence and sequences in a sequence database, wherein the molecular masses of molecules in  
8 the *de novo* sequence are compared to molecular masses of molecules in each sequence in the sequence  
9 database, and

10 c. executes a third program that interprets mass differences of modification sites between the  
11 sequence in the sequence database and the *de novo* sequence that have been identified by the mass-based  
12 alignment as modifications identified in a modification catalog, and

13 d. executes a fourth program that calculates at least one match score for the mass-based  
14 alignment that provides an indication of matching between the sequence in the sequence database and the  
15 *de novo* sequence.

1 89. The computer based system of claim 88, wherein at least a first processor executes one or  
2 more programs that identifies sequences in the sequence database from mass-based alignments in  
3 response to the match scores.

1 90. The computer based system of claim 89, wherein at least a first processor executes one or  
2 more programs that groups identifications of sequences in the sequence database from at least one *de*  
3 *nov*o sequence into an identified macromolecule list that agrees with the *de novo* sequencing results.